Scalable emulation of protein equilibrium ensembles with generative deep learning

Year: 2025 | Citations: 125 | Authors: Sarah Lewis, Tim Hempel, José Jiménez-Luna, Michael Gastegger, Yu Xie

---

Abstract

Following the sequence and structure revolutions, predicting the dynamical mechanisms of proteins that implement biological function remains an outstanding scientific challenge. Several experimental techniques and molecular dynamics (MD) simulations can, in principle, determine conformational states, binding configurations and their probabilities, but suffer from low throughput. Here we develop a Biomolecular Emulator (BioEmu), a generative deep learning system that can generate thousands of statistically independent samples from the protein structure ensemble per hour on a single graphical processing unit. By leveraging novel training methods and vast data of protein structures, over 200 milliseconds of MD simulation, and experimental protein stabilities, BioEmu's protein ensembles represent equilibrium in a range of challenging and practically relevant metrics. Qualitatively, BioEmu samples many functionally relevant conformational changes, ranging from formation of cryptic pockets, over unfolding of specific protein regions, to large-scale domain rearrangements. Quantitatively, BioEmu samples protein conformations with relative free energy errors around 1 kcal/mol, as validated against millisecond-timescale MD simulation and experimentally-measured protein stabilities. By simultaneously emulating structural ensembles and thermodynamic properties, BioEmu reveals mechanistic insights, such as the causes for fold destabilization of mutants, and can efficiently provide experimentally-testable hypotheses.